Appl. Math. J. Chinese Univ.
2021, 36(1): 60-69

Heteroscedastic Laplace mixture of experts regression models and applications

WU Liu-cang[1,*] ZHANG Shu-yu[2] LI Shuang-shuang[3]

Abstract. Mixture of Experts (MoE) regression models are widely studied in statistics and machine learning for modeling heterogeneity in data for regression, clustering and classification. Laplace distribution is one of the most important statistical tools to analyze thick and tail data. Laplace Mixture of Linear Experts (LMoLE) regression models are based on the Laplace distribution which is more robust. Similar to modelling variance parameter in a homogeneous population, we propose and study a new novel class of models: heteroscedastic Laplace mixture of experts regression models to analyze the heteroscedastic data coming from a heterogeneous population in this paper. The issues of maximum likelihood estimation are addressed. In particular, Minorization-Maximization (MM) algorithm for estimating the regression parameters is developed. Properties of the estimators of the regression coefficients are evaluated through Monte Carlo simulations. Results from the analysis of two real data sets are presented.

§1 Introduction

MoE regression models have received considerable attention in various applications and are known as powerful tools in heterogeneous populations. MoE regression models are widely studied in statistics, econometrics and machine learning for modeling heterogeneity in data for regression, clustering and classification. They were first introduced by Jacobs *et al.*[1], included mixing proportions (known as the gating network), and the component densities. A complete review of the MoE regression models can be found in Yuksel *et al.*[2]. MoE regression models for continuous data are usually based on the normal distribution. However, it is well-known that the normal distribution is sensitive to outliers. Recently, Chamroukhi[3] applied mixture of experts based on t distribution to model non-linear regression data. Laplace mixture of experts (LMoE) for non-linear regression data were put forward by Nguyen and McLachlan [4].

Received: 2018-01-15. Revised: 2019-12-16.

MR Subject Classification: 62F10, 62H12.

Keywords: mixture of experts regression models, heteroscedastic mixture of experts regression models, Laplace distribution, MM algorithm.

Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-021-3591-2.

Supported by the National Natural Science Foundation of China (11861041, 11261025).

[*] Corresponding author.

Similar to the ordinary regression models in a homogeneous population, the homoscedasticity of every the component densities (known as the experts) is a basic assumption in the MoE regression model. Under this assumption, it can be feasible to make routine statistical inference. If the variances of observations in the every subpopulations are heterogeneous and unknown, the regression analysis will meet many troubles. Moreover, we encounter that there are many heteroscedastic data around our real life. Therefore, the assumption of the homoscedasticity in the every subpopulations is not consistent with the reality.

To the best of our knowledge, in the framework of the MoE regression models, it is assumed that the equal variance for each component is constant across observations, e.g., Yuksel *et al.* [2] and references therein . However, little work has been done to model variance in the MoE. Huang and Yao [5] investigated models which allow the mixing proportions to depend on the covariates nonparametric. Huang *et al.*[6] proposed a fully nonparametric mixture of regression models by assuming that the mixing proportions, the regression functions, and the variance functions are nonparametric functions of a covariate.

The main objective of this paper is to develop a heteroscedastic mixture of experts regression models determine which variables how to drive the mean and variance parameters in different subpopulations, that is, the mean and variance parameters may change with different covariates in different subgroups of observations. Similar to modelling variance parameter in a homogeneous population (see Aitkin[7]; Taylor and Verbyla[8]; Wu *et al.*[9] and references therein), we propose and study a new novel class of models: heteroscedastic mixture of experts regression models based on the Laplace distribution to analyze the heteroscedastic data coming from a heterogeneous population in this paper. We extend the homogeneous heteroscedasticity data to heterogeneous heteroscedasticity data in Laplace mixture of linear experts. Firstly, we propose the heteroscedastic Laplace mixture of experts regression models. Next, we show the maximum likelihood estimator (MLE) using Minorization-Maximization (MM) algorithm estimates the regression parameters. Properties of the estimators of the regression coefficients are evaluated through Monte Carlo experiments. Results from the analysis of two real data sets are presented.

The paper is organized as follows. The heteroscedastic Laplace mixture of experts regression models are described in Section 2. The MM algorithm for model fitting is given in Section 3. In Section 4, several simulations present the performance of parameter estimation, and Section 5 demonstrates two real data applications of the model. Finally, we conclude with a discussion in Section 6.

§2 Heteroscedastic Laplace mixture of experts regression models

2.1 Heteroscedastic Laplace mixture of experts regression models

The MoE framework can be define as follows. Let $z \in \{1, ..., m\}$ be a categorical random variable such that

$$P(z=j|\nu_i)=\pi_j(\nu_i;\alpha)=\begin{cases}\frac{\exp(\nu_i^T\alpha_j)}{1+\sum\limits_{j'=1}^{m-1}\exp(\nu_i^T\alpha_{j'})}, & if \quad j=1,...,m-1,\\ \frac{1}{1+\sum\limits_{j'=1}^{m-1}\exp(\nu_i^T\alpha_{j'})}, & otherwise.\end{cases} \tag{1}$$

and Figure 1 shows the structure of an MoE regression models with $m=2$ experts, where the two experts are mixed by the gating network.

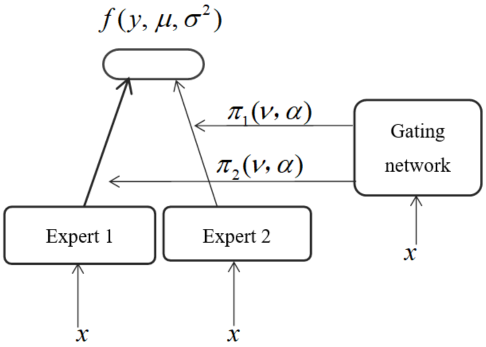

Figure 1. The structure of a MoE model with $m=2$ experts.

Let Y be a response variable that follows Laplace distribution and composed of m categories in an LMoE model, the density function of Y is

$$Laplace_Y(y|\nu,x,h)=\sum_{j=1}^{m}\pi_j(\nu,\alpha)Laplace(y;\mu_j(x,\beta),\sigma_j^2(h,\gamma)). \tag{2}$$

This paper aims at mixture of regression and heteroscedasticity data in Laplace mixture of linear experts, simultaneously modelling mixing proportions, location parameters and scale parameters. We consider the following heteroscedastic Laplace mixture of experts regression models:

$$\begin{cases} y_i & \sim & \sum\limits_{j=1}^{m}\pi_j Laplace(y_i;\mu_{ij},\sigma_{ij}^2),\\ \mu_{ij} & = & x_i^T\beta_j,\\ \log\sigma_{ij}^2 & = & h_i^T\gamma_j,\\ i & = & 1,\dots,n; j=1,\dots,m.\end{cases} \tag{2.3}$$

where:

$$\pi_j=\begin{cases}\frac{\exp(\nu_i^T\alpha_j)}{1+\sum\limits_{j'=1}^{m-1}\exp(\nu_i^T\alpha_{j'})}, & if \quad j=1,...,m-1,\\ \frac{1}{1+\sum\limits_{j'=1}^{m-1}\exp(\nu_i^T\alpha_{j'})}, & otherwise.\end{cases}$$

$\nu_i=(\nu_{i1},...,\nu_{it})^T$, $x_i=(x_{i1},...,x_{ip})^T$, $h_i=(h_{i1},...,h_{iq})^T$ is explain vector and $\alpha_j=(\alpha_{j1},...,\alpha_{jt})^T$, $\beta_j=(\beta_{j1},...,\beta_{jp})^T$, $\gamma_j=(\gamma_{j1},...,\gamma_{jq})^T$ is unknown parameters and $0<\pi_j<1$, $\sum\limits_{j=1}^{m}\pi_j=1$ and $Laplace(y_i;\mu_j,\sigma_j^2)=\frac{1}{\sqrt{2}\sigma_j}\exp(-\frac{\sqrt{2}|y_i-\mu_j|}{\sigma_j})$.

2.2 Identifiability

Identifiability is not a negligible issue in finite mixture models. Titterington *et al.* [11] have given the related conclusion that the finite mixture models of continuous distribution in most cases is identifiable. In this paper, the necessary and sufficient condition that model can be identifiable is:

$$\sum_{j=1}^{m}\pi_j(\nu,\alpha)Laplace(y;\mu_j(x,\beta),\sigma_j^2(h,\gamma))=\sum_{j=1}^{m^*}\pi_j(\nu,\alpha^*)Laplace(y;\mu_j(x,\beta^*),\sigma_j^2(h,\gamma^*)).$$

If and only if $m=m^*$, $\alpha=\alpha^*$, $\beta=\beta^*$,$\gamma=\gamma^*$. Obviously the model is identifiable, for different parameters, the corresponding distribution is also different, up to a permutation.

§3 Parameter Estimation

3.1 α phase

EM algorithm require specialist knowledge of probabilistic characterizations in order to express the iterative updates. Instead of EM algorithm, we will introduce a equation (4) that proposed by Nguyen and McLachlan [4], see Section 2.1 of Nguyen and McLachlan [4], a monotonic iterative scheme using the Minorization-Maximization (MM) algorithm (Hunter and Lange [10]) framework, which can violate the usual monotonicity to make the update:

$$\alpha_j^{(k+1)}=4\varDelta^{-1}\delta_j^{(k+1)}+\delta_j^{(k)},\tag{1}$$

where : $\varDelta=\sum\limits_{j=1}^{n}\nu_j\nu_j^T$, $\delta_j^{(k+1)}=\sum\limits_{j=1}^{n}[\tau_{ij}^{(k+1)}-\pi_j(\nu_i;\alpha_j^{(k)})]\nu_i$,

$$\tau_{ij}^{(k+1)}=\frac{\pi_j(\nu_i;\alpha^{(k)})Laplace(y_i;x_i^T\beta_j^{(k)},h_i^T\gamma_j^{(k)})}{\sum\limits_{j=1}^{m}\pi_j(\nu_i;\alpha^{(k)})Laplace(y_i;x_i^T\beta_j^{(k)},h_i^T\gamma_j^{(k)})}.$$

3.2 β and γ phase

First, we should determine the number of components. As BIC (Bayesian information criterion) tends to outperform the other criteria, such as, AIC (Akaike information criterion), CLC (Classification likelihood criterion), it can be applied this section(McLachlan and Peel[12]). Suppose that $m_0\in\{r_1,...,r_s\}$ is the true value of m. For each , we fit heteroscedastic Laplace mixture of experts regression models with r_l components and compute its MLE $\hat\theta_{(l)n}$. The BIC for each l can be given as

$$BIC(l)=-2\log L_n(\hat\theta_{(l)n})+\log n[r_l(t+p+q)-t],\tag{2}$$

and $r_l(t+p+q)-t$ is the total number of parameter elements in model l. Under BIC criteria the number of components is selected by $m=r_{\hat l}$,using the rule $\hat l=\arg\min BIC(l)$.

Then, EM algorithm (McLachlan and Krishnan[13]) is utilized to estimate β and γ phase. The flow path divided into two steps: E-step and M-step. E-step calculates the expectation of logarithmic likelihood function according to the parameters initial values or the result of the previous iteration. M-step maximizes the logarithm likelihood function to get the new parameter values. Using the new parameter values to instead of initial values or the previous iteration results, repeat the above two steps until convergence.

The specific procedures of EM algorithms:

E-step: Utilize the $\theta^{(k)}$ and $\alpha^{(k)}$ to calculate $\hat{\tau}_{ij}^{(k+1)}$:

$$\hat{\tau}_{ij}^{(k+1)} = E(\tau_{ij}^{(k)}|y_i, x_i, h_i, \theta^{(k)}, \alpha^{(k)}) = \frac{\pi_j(\nu_i; \alpha^{(k)}) Laplace(y_i; x_i^T \beta_j^{(k)}, h_i^T \gamma_j^{(k)})}{\sum\limits_{j=1}^{m} \pi_j(\nu_i; \alpha^{(k)}) Laplace(y_i; x_i^T \beta_j^{(k)}, h_i^T \gamma_j^{(k)})}.$$

Then calculate expectation of logarithmic likelihood function:

$$Q(\cdot) = E(l_c|y_i, x_i, h_i, \theta^{(k)}, \alpha^{(k)}).$$

M-step: Using the Gaussian–Newton algorithm for the simultaneous maximum likelihood estimate of $\Theta_j = (\beta_j^T, \gamma_j^T)^T$, the the updated estimates at the (k+1)th iteration are

$$\Theta_j^{(k+1)} = \Theta_j^{(k)} + [-\frac{\partial^2 Q}{\partial \Theta_j \partial \Theta_j^T}(\Theta_j^{(k)})]^{-1} U(\Theta_j^{(k)}), j = 1, 2, \cdots, m,$$

where $\frac{\partial^2 Q_2}{\partial \Theta_j \partial \Theta_j^T}(\Theta_j)$ is the observed Fisher information matrix and $U(\Theta_j)$ is the score function. The E-step and M-step are alternated repeatedly until convergence is obtained.

See Appendix for detailed calculation process.

§4 Monte Carlo Simulation

To evaluate the performance of the proposed MM and EM estimation algorithms, we conduct some Monte Carlo simulations. The performance of estimator $\hat{\theta}$, will be assessed by using the mean square error (MSE), defined as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta_0)^T(\hat{\theta}) = E(\hat{\theta} - \theta_0).$$

We simulate data from the following model (4.1)

$$\begin{cases} y_i & \sim & \pi_1 Laplace(y_i; \mu_{i1}, \sigma_{i1}^2) + \pi_2 Laplace(y_i; \mu_{i2}, \sigma_{i2}^2), \\ \mu_{ij} & = & x_i^T \beta_j, \\ \log \sigma_{ij}^2 & = & h_i^T \gamma_j, \\ i & = & 1, \ldots, n; j = 1, 2. \end{cases} \tag{4.1}$$

where:

$$\pi_j = \begin{cases} \frac{\exp(\nu_i^T \alpha_j)}{1+\exp(\nu_i^T \alpha_j)}, & j = 1, \\ \frac{1}{1+\exp(\nu_i^T \alpha_j)}, & j = 2. \end{cases}$$

To perform this simulation, we take the $\alpha_1 = (0, 1)^T$, $\beta_1 = (0, 1)^T$, $\beta_2 = (0, -1)^T$, $\gamma_1 = (0, 1)^T$, $\gamma_2 = (0, -1)^T$, and $\nu_j \sim U(-1, 1)$, $x_j \sim U(-1, 1)$ and $h_j \sim U(-1, 1)$. y_i is generated according to the model (4.1). The sample sizes considered are $n = 64, 128, 256, 512, 1024$. The following simulation results are all based on 1000 independent repetitions. Table 1 reports the average of estimator. Furthermore, the column labeled “Mean” and “MSE” give the average estimators and the mean square errors of $\hat{\beta}_j$, $\hat{\gamma}_j$, $j = 1, 2$ and $\hat{\alpha}_1$.

Table 1. Simulation results for the model (4.1).

Parameter	Sample Size	Mean	MSE
α_1	64	$(-0.0078, 0.9646)^T$	0.1126
	128	$(-0.0000, 0.9597)^T$	0.0524
	256	$(0.0005, 0.9498)^T$	0.0285
	512	$(-0.0000, 0.9507)^T$	0.0150
	1024	$(-0.0018, 0.9572)^T$	0.0083
β_1	64	$(0.0140, 1.0025)^T$	0.4173
	128	$(0.0121, 1.0016)^T$	0.1871
	256	$(0.0020, 1.0090)^T$	0.0867
	512	$(-0.0017, 1.0013)^T$	0.0397
	1024	$(0.0020, 1.0053)^T$	0.0211
γ_1	64	$(-0.0223, 0.9204)^T$	0.5747
	128	$(0.0079, 0.9657)^T$	0.3311
	256	$(0.0081, 0.9762)^T$	0.1929
	512	$(-0.0038, 0.9879)^T$	0.1039
	1024	$(-0.0051, 0.9967)^T$	0.0521
β_2	64	$(0.0029, -0.9975)^T$	0.4267
	128	$(-0.0033, -1.0075)^T$	0.1847
	256	$(0.0007, -0.9787)^T$	0.0889
	512	$(-0.0035, -0.9882)^T$	0.0411
	1024	$(0.0003, -1.0039)^T$	0.0202
γ_2	64	$(-0.0060, -0.9150)^T$	0.5545
	128	$(0.0147, -0.9559)^T$	0.3283
	256	$(0.0022, -0.9890)^T$	0.1831
	512	$(-0.0084, -0.9789)^T$	0.1026
	1024	$(-0.0042, -1.0079)^T$	0.0510

Table 1 shows that the following observations:

(1) The results in Table 1 indicate that the MM algorithm performs well in estimating the coefficients. The MSEs for all the parameters decrease as the sample size increases from 64 to 1024.

(2) For the given sample size n, the performance of maximum likelihood estimation in the location model is significantly better than that of maximum likelihood estimation in the scale model in terms of the MSE.

§5 Application

In this section, two real data sets from the Air Quality Index(AQI) data[14] and Sheep Time Series Data[15] are used to illustrate the proposed methods.

5.1 Air Quality Index

In recent years, Beijing as the capital of China, is also China's political and cultural center, air pollution issues are quite serious, which must be pay much attention, and to do a good job of environmental protection in Beijing is great significance. Air Quality Index (AQI) is a quantitative description of Air Quality Index dimensionless. The main pollutants is divided into six types: fine particulate matter, particulate matter, sulfur dioxide, nitrogen dioxide, ozone, carbon monoxide.

We collected 365 days AQI data of Beijing in 2015 from Ministry of Environmental Protection of the Peoples Republic of China Data Center(*http://datacenter.mep.gov.cn/*)[14], explanatory variables are X_1–fine particulate matter $(PM2.5, \mu g/m^3)$, X_2–particulate matter $(PM10, \mu g/m^3)$, X_3–sulfur dioxide $(SO_2, \mu g/m^3)$, X_4–nitrogen dioxide $(NO_2, \mu g/m^3)$, X_5–ozone $(O_3, \mu g/m^3)$, X_6–carbon monoxide $(CO, mg/m^3)$, respectively. According to the results of cluster analysis, we get the BIC values: BIC(2)=4504.903 and BIC(3)=4574.886, so we choose to use two-component heteroscedastic Laplace mixture of experts regression models for fitting the data.

The data can be divided into two categories by contrast observing, heavier pollution area (including January, February, March, April, October, November, December seven months) and lighter pollution area (including May, June, July, August, September five months).

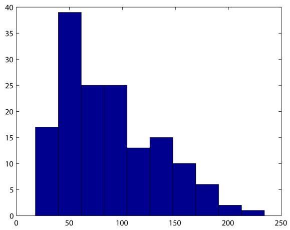

Figure 2. Histogram of heavier area.

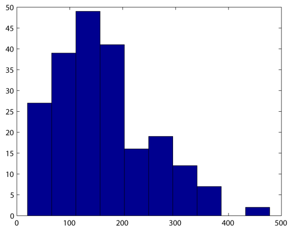

Figure 3. Histogram of lighter area.

Figure 2 and Figure 3 show the histogram of AQI data for heavier pollution area and lighter pollution area. We find the histogram of the two parts of AQI data is different obviously, need to use mixture models to fit. Applying the proposed model (2.3), MLE estimators for the parameters see Table 2.

Table 2. The results for the Air Quality Index(AQI) data.

Parameter	Const	X_1	X_2	X_3	X_4	X_5	X_6
α_1	5.540	0.022	0.027	0.321	-0.056	-2.938	-0.098
β_1	63.226	1.095	-0.029	0.315	-0.035	-11.708	-0.124
β_2	7.239	0.980	0.140	-0.342	0.012	0.639	0.214
γ_1	9.716	-0.016	0.016	-0.046	-0.041	-0.107	-0.014
γ_2	0.972	0.004	-0.001	-0.038	0.002	-0.061	0.019

In fact, the AQI values of Beijing city are influenced by season and month in different degrees, this is because Beijing is located in the north of China. From November each year to March next year, Beijing city need heating to residents, then chemical compounds make contribution to pollute the air produce by burning coal.

5.2 Sheep Time Series Data

In addition, we also found that heteroscedastic Laplace mixture of experts regression models can fit time series data better, we collected data, annual sheep population in England and Wales 1867–1939 from Time Series Data Library (*http://robjhyndman.com/TSDL/*)[15]. Let the dependent y is the sheep population, explaining variables for $\nu_j = x_j = h_j = (1, t_j)$, $t_j = j + 1867$, here $j = 1, ..., 73$.

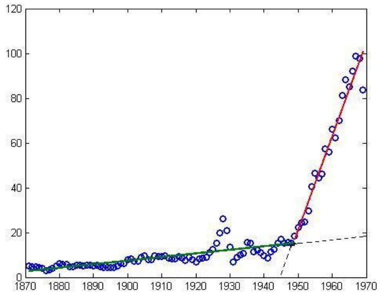

Figure 4. The scatter of sheep population.

Observing the scatter from Figure 4, we consider two-component heteroscedastic Laplace mixture of experts regression models should be used to fitting, and BIC values confirmed our speculation: BIC(2)=43.239 and BIC(3)=47.720. Applying the proposed model (2.3), MLE estimators for the parameters see Table 3.

Table 3. The results for the Sheep Time Series Data.

Parameters	Const	X_1
α_1	42491.956	-21.796
β_1	-222.185	0.121
β_2	-8375.5	0.004
γ_1	-70.096	0.037
γ_2	-278.421	0.143

We find that the growth speed of sheep number has changed a lot around 1950, we refer to the relevant information, that may be IVF (In Vitro Fertilization) technique succeed in the 1950s, and developed rapidly in next 20 years, now become an important guide and conventional animal breeding biotechnology in developed countries such as Europe, America and Oceania.

§6 Conclusion

In this paper, in order to analyze the heteroscedastic data coming from a heterogeneous population, we proposed and studied a new novel class of models: heteroscedastic Laplace mixture of experts regression models. We developed a Minorization-Maximization (MM) algorithm to estimate the regression parameters. The obtained results on simulated data show the good performance. The analysis and applications of two practical data confirm the usefulness for the heteroscedastic Laplace mixture of experts regression models.

As a final remark, we only considered the heteroscedastic Laplace mixture of experts regression models in their standard (non-hierarchical) version. One interesting future direction is therefore to extend the proposed models to the hierarchical MoE framework (Jordan and Jacobs[16]). Furthermore, a natural future extension of this work is to consider the case of heteroscedastic Laplace mixture of experts regression models for multiple regression on multivariate data rather than simple regression on univariate data.

Appendix

In this section, we add some derivatives pertinent in M-step of EM algorithms.

$$\frac{\partial^2 Q}{\partial\Theta_j\partial\Theta_j^T} = \begin{bmatrix} \frac{\partial^2 Q}{\partial\beta_j\partial\beta_j^T} & \frac{\partial^2 Q}{\partial\beta_j\partial\gamma_j^T} \\ \frac{\partial^2 Q}{\partial\gamma_j\partial\beta_j^T} & \frac{\partial^2 Q}{\partial\gamma_j\partial\gamma_j^T} \end{bmatrix},$$

$$\frac{\partial^2 Q}{\partial\beta_j\partial\beta_j^T} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}(\frac{\sqrt{2}}{\sigma_{ij}}\cdot\frac{\partial^2\mu_{ij}}{\partial\beta_j\partial\beta_j^T}),$$

$$\frac{\partial^2 Q}{\partial\gamma_j\partial\gamma_j^T} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}[(-\frac{1}{\sigma_{ij}^2}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j^T}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j} - \frac{1}{\sigma_{ij}}\cdot\frac{\partial^2\sigma_{ij}}{\partial\gamma_j\partial\gamma_j^T}) + (\frac{2\sqrt{2}\mid y_i-\mu_{ij}\mid}{\sigma_{ij}^3}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j^T}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j} - \frac{\sqrt{2}\mid y_i-\mu_{ij}\mid}{\sigma_{ij}^2}\cdot\frac{\partial^2\sigma_{ij}}{\partial\gamma_j\partial\gamma_j^T})]$$

$$\frac{\partial^2 Q}{\partial\beta_j\partial\gamma_j^T} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}(-\frac{\sqrt{2}}{\sigma_{ij}^2}\cdot\frac{\partial\mu_{ij}}{\partial\beta_j}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j^T}),$$

$$\frac{\partial^2 Q}{\partial\gamma_j\partial\beta_j^T} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}(\frac{\sqrt{2}}{\sigma_{ij}^2}\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j}\cdot\frac{\partial\mu_{ij}}{\partial\beta_j^T}),$$

$$\frac{\partial Q}{\partial\beta_j} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}(\frac{\sqrt{2}}{\sigma_{ij}}\cdot\frac{\partial\mu_{ij}}{\partial\beta_j}),$$

$$\frac{\partial Q}{\partial\gamma_j} = \sum_{i=1}^n \hat{\tau}_{ij}^{(k+1)}(-\frac{1}{\sigma_{ij}} - \frac{\sqrt{2}\mid y_i-\mu_{ij}\mid}{\sigma_{ij}^2})\cdot\frac{\partial\sigma_{ij}}{\partial\gamma_j}.$$

Here referred a method of taking derivative to absolute about parameter β_j, we can consider a function $f(\beta_j) = \sum\limits_{i=1}^n \mid y_i - x_i^T\beta_j \mid$ derivative to β_j: $\frac{\partial f(\beta)}{\partial\beta_j} = -\sum\limits_{i=1}^n x_i sgn(y_i - x_i^T\beta_j)$, where $sgn(\cdot)$ is the sign function which takes -1, 0, 1 if the argument is negative, 0, and positive respectively. Let $\omega_i = \frac{1}{|y_i - x_i^T\beta_j|}$, and $\frac{\partial f(\beta_j)}{\partial\beta_j} = \sum\limits_{i=1}^n \omega_i x_i sgn(y_i - x_i^T\beta_j)$. Thus we can defuse this problem.

References

[1] R A Jacobs, M I Jordan, S J Nowlan, G E Hinton. *Adaptive mixtures of local experts*, Neural Computation, 1991, 3: 79-87.

[2] S E Yuksel, J N Wilson, P D Gader. *Twenty years of mixture of experts*, IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(8): 1177-1193.

[3] F Chamroukhi. *Robust mixture of experts modeling using the t distribution*, Neural networks, 2016, 79: 20-36.

[4] H D Nguyen, G J McLachlan. *Laplace mixture of linear experts*, Computational statistics and Data analysis, 2016, 93: 177-191.

[5] M Huang, W X Yao. *Mixture of regression models with varying mixing proportions: a semiparametric approach*, Journal of the American Statistical Association, 2012, 107: 711-724.

[6] M Huang, R Z Li, S L Wang. *Nonparametric mixture of regression models*, Journal of the American Statistical Association, 2013, 108: 929-941.

[7] M Aitkin. *Modelling variance heterogeneity in normal regression using GLIM*, Journal of the Royal Statistical Society, Series C, 1987, 36: 332-339.

[8] J T Taylor, A P Verbyla. *Joint modelling of location and scale parameters of the t distribution*, Statistical Modelling, 2004, 4: 91-112.

[9] L C Wu, Z Z Zhang, D K Xu. *Variable selection in joint location and scale models of the skew-normal distribution*, Journal of Statistical Computation and Simulation, 2013, 83(7): 1266-1278.

[10] D R Hunter, K Lange. *A tutorial on MM algorithms*, The American Statistician, 2004, 58(1): 30-37.

[11] D M Titterington, A F M Smith, U E Makov. *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.

[12] G J McLachlan, D Peel. *Finite Mixture Models*, Wiley, New York, 2000.

[13] G J McLachlan, T Krishnan. *The EM Algorithm and Extensions*, Wiley, New York, 2008.

[14] Ministry of Environmental Protection of the Peoples Republic of China Data Center, *http://datacenter.mep.gov.cn/*.

[15] Time Series Data Library, *http://robjhyndman.com/TSDL/*.

[16] M I Jordan, R A Jacobs. *Hierachical mixtures of experts and the EM algorithm*, Neural Computation, 1994, 6: 181-214.

[1,2]Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China.
Email: wuliucang@163.com, 365669601@qq.com

[3]College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China.
Email: 1028273976@qq.com